Approximate Bayesian Inference for Geostatistical Generalised Linear Models

Evangelos Evangelou

Department of Mathematical Sciences – University of Bath, BA2 7AY, Bath, UK

ee224@bath.ac.uk

Abstract: The aim of this paper is to bring together recent developments in Bayesian generalised linear mixed models and geostatistics. We focus on approximate methods on both areas. A technique known as full-scale approximation, proposed by Sang and Huang (2012) for improving the computational drawbacks of large geostatistical data, is incorporated into the INLA methodology, used for approximate Bayesian inference. We also discuss how INLA can be used for approximating the posterior distribution of transformations of parameters, useful for practical applications. Issues regarding the choice of the parameters of the approximation such as the knots and taper range are also addressed. Emphasis is given in applications in the context of disease mapping by illustrating the methodology for modelling the *loa loa* prevalence in Cameroon and malaria in the Gambia.

Keywords: Disease mapping; Full-scale approximation; Generalised linear spatial model; Geostatistics; Integrated nested Laplace approximation.

1 Introduction

Since the paper by Diggle et al. (1998), Bayesian methods for fitting spatial geostatistical models have become the norm. The models are fitted using Markov chain Monte-Carlo methods, or other simulation-based methods (e.g Zhang, 2004), but with the drawback of being too computationally intensive in order to obtain good samples. Furthermore, large amounts of geostatistical data are impossible to handle with the common computer. Recently, a number of papers have appeared in the literature that aim to overcome these computational disadvantages of Bayesian spatial models. This

Date: February 11, 2019

paper aims to illustrate the application of some of these methods to the geostatistical generalised linear model (GGLM) (the term "geostatistical generalised linear model", as opposed to the more general "spatial generalised linear model", is used here to emphasise the geostatistical feature of our model); as well as to address issues related to the implementation of these methods.

Rue et al. (2009) introduced the method of integrated, nested, Laplace approximation (INLA) for fitting Bayesian mixed models. The general setting assumes that one can write down the conditional likelihood of the observations given the value of the random effect. The random effect is assumed to be Gaussian with variance depending on a small number of parameters. Moreover, the fixed effects are assumed to have a Gaussian or uniform prior. Inference is performed by approximating the unconditional distribution of the observations using Laplace's method (Barndorff-Nielsen and Cox, 1989) and emphasis is given to the case where the variance-covariance matrix of the random effects has a Markov random field structure (Rue and Held, 2005). Despite some criticism (Taylor and Diggle, 2014), the method has been shown to be relatively accurate in several settings (Eidsvik et al., 2009, Paul et al., 2010, Martino et al., 2011, Schrödle and Held, 2011, Illian et al., 2012).

One advantage of the Markov random field structure, compared to the geostatistical random field, is computational. In the former the precision matrix can be written explicitly in terms of the variance parameters and is sparse. Therefore, calculations involving the Gaussian likelihood are performed in short computational time using routines for sparse matrices. On the other hand, spatial data over a continuous area are naturally represented by a geostatistical random field.

Nevertheless, the assumptions made by INLA are satisfied for geostatistical models and, to that end, Eidsvik et al. (2009) illustrated its application successfully to the GGLM. One point that was only briefly discussed in their paper is the application of the method when the number of sampling locations is large. In this case, the inverse of a large dense matrix is required at every iteration of the Laplace approximation which can have detrimental effects in the speed and accuracy of the method. In a subsequent paper, Eidsvik et al. (2012) account for this by considering the predictive process approximation to the covariance (Banerjee et al., 2008) but found that the approximation is sensitive to the number of knots.

Motivated by the fast calculations for sparse matrices and the success of INLA in the Markov random field case, Lindgren et al. (2011) developed a theory for approximating spatial random

fields by Markov random fields and took advantage of the computational benefits of the latter. This method has been illustrated in subsequent papers by Simpson et al. (2012) and Cameletti et al. (2013) among others. On the other hand, this method applies only for certain covariance functions, namely a subset of the Matérn family.

An alternative approach, termed *full scale approximation*, has been suggested by Sang and Huang (2012) by building on the idea of the predictive process model, combined with covariance tapering (Furrer et al., 2006). The idea is to replace the dense covariance matrix of the random field, Σ say, by a matrix of the form $\Sigma_S + \Gamma\Sigma_L^{-1}\Gamma^\intercal$ where Σ_S is sparse and Σ_L is of low dimension. The benefit is that the new matrix is easier to handle computationally and is a good approximation to the original matrix. Sang et al. (2011) and Sang and Huang (2012) apply this approximation to Gaussian models. An interesting question answered in this paper is how we can incorporate this approximation into INLA.

There are cases, e.g. in the context of disease mapping, where interest lies in predicting a non-linear transformation of the spatial random effects. For MCMC this is straightforward by applying the transformation to the MCMC sample but for INLA, unless the distribution of the transformed variable is known in closed form, it is not so clear. This question is answered in the present paper by first approximating the predictive distribution of the spatial random effects by a mixture of normal distributions, and then use that to approximate the mean of the transformation via a weighted average.

The remaining of the paper is organised as follows. The GGLM is introduced in Section 2. Section 3 describes the INLA methodology and discusses its application to GGLM. Three examples are presented in Section 4 illustrating the techniques proposed. Finally, Section 5 presents the conclusions of this article. Some technical details are put in the Appendix.

2 Model formulation

We assume a continuous domain of interest $\mathbb{S}$ over which a Gaussian random field $\mathcal{Z}$ is defined. That is, for every finite subset $S = \{s_1, \ldots, s_k\} \subset \mathbb{S}$, the value of $\mathcal{Z}$ on S follows the k-dimensional normal distribution with mean 0 and variance-covariance matrix Σ. The (i, j) element of the $k \times k$

matrix Σ has the form

$$\sigma_{ij} = \sigma^2 \, r(d_{ij}; \rho),$$

where $\sigma^2 > 0$ is the so-called *sill* parameter, interpreted as the variance of the random field at each location, r is a known function of the distance d_{ij} between locations s_i and s_j called the *correlogram*, giving the correlation between the components of the random field having distance d_{ij} apart, and $\rho > 0$ is parameter of the correlogram called the *range*. Two forms of the correlogram that we will use in the examples of Section 4 are

- *Exponential*: $r(d; \rho) = \exp(-d/\rho)$,
- *Spherical*: $r(d; \rho) = 1 - 1.5\,(d/\rho) + 0.5\,(d/\rho)^3$, if $d < \rho$; and 0 otherwise.

We assume that a total of N observations $\boldsymbol{y} = \{y_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i\}$ are taken within the spatial domain $\mathbb{S}$ with the (i, j)th observation corresponding to location $s_i \in \mathbb{S}$, and n_i denotes the number of observations at location i with $\sum n_i = N$. The n_i's may, for example denote the number individuals sampled or the number of observations from a single individual. Finally, each y_{ij} corresponds to the total of m_{ij} replications of the (i, j) experiment, e.g. for binary data m_{ij} corresponds to the number of trials of the experiment and for Poisson data to the length of unit time that the sampling is taking place.

We denote by S the collection of the k sampled locations, and by $\boldsymbol{z}$ the value of $\mathcal{Z}$ on S. The observations are modelled by a GGLM, that is, y_{ij} is independent of $y_{i'j'}$ for $i \neq i'$ or $j \neq j'$ conditioned on the linear predictor $\boldsymbol{w}$, with distribution from the exponential family.

Notable members of the exponential family include the binomial and Poisson distributions. The binomial model with logit link has distribution

$$f(y|w) = \exp\left\{y\, w - m \log(1 + e^w)\right\} \binom{m}{y},$$

and the Poisson model with logarithmic link is

$$f(y|w) = \exp\left\{y\, w - m \exp(w)\right\} / y!.$$

These distributions have the general form

$$f(y|w) = \exp\{y\, w - m\kappa(w)\}\, a(y),$$

where $\kappa(\cdot)$ is known as the *cumulant function* (McCullagh and Nelder, 1999).

The (i,j) component of the N-dimensional linear predictor $\boldsymbol{w}$, w_{ij}, is assumed to have the form

$$w_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + z_i + \epsilon_{ij},\ i = 1,\ldots,k,\ j = 1,\ldots,n_i,$$

where $\boldsymbol{x}_{ij}$ denotes a vector of p explanatory variables corresponding to the jth sample at location s_i, β corresponds to a p-dimensional vector of regressor coefficients, z_i denotes the value of the Gaussian random field at location s_i and ϵ_{ij} is a zero-mean random component, independent of the z_i's, corresponding to other imponderable effects associated with the (i,j)th individual.

Let $\boldsymbol{\epsilon}$ denote the N dimensional random vector of ϵ_{ij}'s ordered by location, i.e. $\boldsymbol{\epsilon} := (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \ldots)$, with variance-covariance matrix D, parametrised by θ_ϵ. An important condition for the implementation of the full-scale approximation discussed in section 3.4 is that the precision matrix, D^{-1}, is sparse. This condition is satisfied, for example, in the i.i.d. case, $D = \tau^2 I$, where I denotes the identity matrix of appropriate dimension, and $\theta_\epsilon = \tau^2 > 0$ is called the *nugget*. Another case, which won't be discussed here but is left as future research, is the Autoregressive model, e.g. the AR(1) model: $\epsilon_{i1} = \delta_{i1}$ and $\epsilon_{ij} = \phi\,\epsilon_{i,j-1} + \delta_{ij}$ for $j = 2,\ldots,n_i$ with $\delta_{ij} \sim N(0,\tau^2)$ i.i.d., is another class of models that falls in the current framework which may be used for incorporating non-spatial correlation within locations (Evangelou and Maroulas, 2017). In this case $\theta_\epsilon = (\tau^2, \phi) \in (0,\infty)\times(-1,1)$ and D^{-1} is block-diagonal of k blocks with each block being a tridiagonal matrix.

Let X be the $N \times p$ matrix with rows the $\boldsymbol{x}_{ij}$'s, and let A be the $N \times k$ matrix of 0's and 1's, with the 1 in each row indicating which component of the random field $\boldsymbol{z}$ is present in the (i,j) element of the linear predictor. In other words we write

$$\boldsymbol{w} = X\beta + A\boldsymbol{z} + \boldsymbol{\epsilon}.$$

Thus the linear predictor $\boldsymbol{w}$ follows the N-dimensional normal distribution with mean $X\beta$ and

variance-covariance matrix

$$\mathsf{T} = A\Sigma A^{\mathsf{T}} + D,$$

Let $\theta = (\sigma^2, \rho, \theta_\epsilon)$ denote the parameters in T, which we will refer to as *variance parameters*.

Our objective is, given the data $\boldsymbol{y}$, to predict the value of the random field at locations $Q = \{q_1, \ldots, q_m\} \subset \mathbb{S}$, and estimate the parameters β, θ. We adopt a Bayesian approach where we assume prior distributions for the parameters, and their posterior distributions conditioned on the data, as well as the conditional distribution of the random field, are sought.

We use the generic symbol $f(\cdot)$ to denote the probability density/mass function of the expression in the parentheses. In the absence of any information about the explanatory variables, it is common to assume flat improper priors for the regressor coefficients β (Christensen et al., 2000, Christensen and Waagepetersen, 2002)

$$f(\beta) \propto 1.$$

Another popular prior considered in the literature is the p-dimensional normal prior with mean say ξ_0 and precision Φ_0, i.e.

$$f(\beta) \propto \exp\left\{-\frac{1}{2}(\beta - \xi_0)^{\mathsf{T}}\Phi_0(\beta - \xi_0)\right\}.$$

For the moment we will not specify any priors for the variance parameters θ but note that for the range parameter, an improper prior leads to an improper posterior (Christensen et al., 2000).

3 Methodology: The INLA approach

3.1 General

Suppose we observe $\boldsymbol{y}$ and we wish to predict the N-dimensional random variable $\boldsymbol{w}$. The distribution of either $\boldsymbol{y}$ or $\boldsymbol{w}$ may depend on an unknown parameter θ. Furthermore, suppose that the distribution of $\boldsymbol{w}|\theta$ is Gaussian. If θ was known, then prediction is performed by evaluating the predictive distribution of $\boldsymbol{w}|\boldsymbol{y};\theta$,

$$f(\boldsymbol{w}|\boldsymbol{y};\theta) \propto f(\boldsymbol{y}, \boldsymbol{w}|\theta), \tag{1}$$

where $f(\boldsymbol{y}, \boldsymbol{w}|\theta) = f(\boldsymbol{y}|\boldsymbol{w}; \theta) f(\boldsymbol{w}|\theta)$. In most applications of generalised linear mixed models the normalising constant

$$f(\boldsymbol{y}|\theta) = \int f(\boldsymbol{y}, \boldsymbol{w}|\theta)\, \mathrm{d}\boldsymbol{w} \tag{2}$$

is not available in closed form and exact evaluation of (1) is impossible. (See for example Breslow and Clayton (1993) and Rue et al. (2009).) The idea of the Gaussian approximation (Rue et al., 2009) is to approximate (1) by a Gaussian density centred around

$$\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}(\theta) = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, f(\boldsymbol{y}, \boldsymbol{w}|\theta),$$

and precision

$$\hat{H} = \hat{H}(\theta) = -\frac{\partial^2}{\partial \boldsymbol{w} \partial \boldsymbol{w}^{\intercal}} \log f(\boldsymbol{y}, \boldsymbol{w}|\theta)\bigg|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}.$$

The values of $\hat{\boldsymbol{w}}$ and $\hat{H}$ are implicitly functions of $(\boldsymbol{y}, \theta)$. Thus, the approximation to (1), obtained by second-order Taylor expansion, is

$$\hat{f}(\boldsymbol{w}|\boldsymbol{y}; \theta) = (2\pi)^{-\frac{N}{2}} |\hat{H}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{w}})^{\intercal} \hat{H} (\boldsymbol{w} - \hat{\boldsymbol{w}})\right\}. \tag{3}$$

Since the distribution of $\boldsymbol{w}|\theta$ is Gaussian, the Gaussian approximation to the predictive distribution is not unreasonable. Alternatively Hosseini et al. (2011) proposed the use of the skew normal as a better approximation when skewness is more apparent and the Gaussian approximation is inaccurate, however this approach will not be pursued in this paper.

The normalising constant (2) is approximated using Laplace approximation (Barndorff-Nielsen and Cox, 1989) by

$$\hat{f}(\boldsymbol{y}|\theta) = (2\pi)^{\frac{N}{2}} |\hat{H}|^{-\frac{1}{2}} f(\boldsymbol{y}, \hat{\boldsymbol{w}}|\theta) = \left.\frac{f(\boldsymbol{y}, \boldsymbol{w}|\theta)}{\hat{f}(\boldsymbol{w}|\boldsymbol{y}; \theta)}\right|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}. \tag{4}$$

Now suppose θ is unknown and assign a prior $\theta \sim f(\theta)$. Then, the posterior density for θ is

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta) f(\theta)}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\theta) f(\theta).$$

From (4) we can approximate

$$\hat{f}(\theta|\boldsymbol{y}) \propto \left.\frac{f(\boldsymbol{y},\boldsymbol{w}|\theta)f(\theta)}{\hat{f}(\boldsymbol{w}|\boldsymbol{y};\theta)}\right|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}. \tag{5}$$

Similarly, for prediction, (1) needs to be integrated with respect to $f(\theta|\boldsymbol{y})\,\mathrm{d}\theta$, i.e. we need to evaluate

$$f(\boldsymbol{w}|\boldsymbol{y}) = \int f(\boldsymbol{w}|\boldsymbol{y};\theta)f(\theta|\boldsymbol{y})\,\mathrm{d}\theta. \tag{6}$$

Using (3) and (5) in (6) we arrive at the following approximation

$$f(\boldsymbol{w}|\boldsymbol{y}) \approx \int \hat{f}(\boldsymbol{w}|\boldsymbol{y};\theta)\hat{f}(\theta|\boldsymbol{y})\,\mathrm{d}\theta. \tag{7}$$

On the other hand, as the distribution $f(\theta)$ is in general non-Gaussian, application of Laplace approximation to the integral in (7) would be inefficient. Instead, we can view (7) as a "continuous mixture" of multivariate normal densities, $\hat{f}(\boldsymbol{w}|\boldsymbol{y};\theta)$, indexed by θ, with weights given by $\hat{f}(\theta|\boldsymbol{y})$. As such it can be approximated by a finite mixture of size a at selected points $\{\theta_1,\ldots,\theta_a\}$ (see Appendix A on how to obtain such values), giving rise to the multivariate normal mixture approximation,

$$\hat{f}(\boldsymbol{w}|\boldsymbol{y}) = \sum_{j=1}^{a} W_j \hat{f}(\boldsymbol{w}|\boldsymbol{y};\theta_j), \text{ where } W_j = \frac{\hat{f}(\theta_j|\boldsymbol{y})}{\sum_k \hat{f}(\theta_k|\boldsymbol{y})}, \tag{8}$$

are the mixture weights.

Using properties of mixture distributions, the mean of (8) is

$$\widehat{\mathrm{E}}(\boldsymbol{w}|\boldsymbol{y}) = \int \boldsymbol{w}\hat{f}(\boldsymbol{w}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{w} = \sum_{j=1}^{a} W_j \int \boldsymbol{w}\hat{f}(\boldsymbol{w}|\boldsymbol{y};\theta_j)\,\mathrm{d}\boldsymbol{w} = \sum_{j=1}^{a} W_j \hat{\boldsymbol{w}}_j, \tag{9}$$

and is the *approximate best predictor* for $\boldsymbol{w}$, with variance, obtained similarly,

$$\widehat{\mathrm{Var}}(\boldsymbol{w}|\boldsymbol{y}) = \sum_{j=1}^{a} W_j \hat{H}_j^{-1} + \sum_{j=1}^{a} W_j \{\hat{\boldsymbol{w}}_j - \hat{\mathrm{E}}(\boldsymbol{w}|\boldsymbol{y})\}\{\hat{\boldsymbol{w}}_j - \hat{\mathrm{E}}(\boldsymbol{w}|\boldsymbol{y})\}^{\intercal}.$$

In the above, we denote by $\hat{\boldsymbol{w}}_j$ and $\hat{H}_j$ the mean and precision respectively of the Gaussian approximation (3) at $\theta = \theta_j$, $j = 1,\ldots,a$.

Let $\boldsymbol{u}$ be another random quantity of dimension n, for example $\boldsymbol{u}$ may represent components

of the random field or the fixed effects, whose distribution is assumed to be independent of $\boldsymbol{y}$ conditioned on $\boldsymbol{w}$, with conditional distribution

$$f(\boldsymbol{u}|\boldsymbol{w};\theta) = (2\pi)^{-\frac{n}{2}}|\Omega|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{u}-\Lambda\boldsymbol{w}-\boldsymbol{\lambda})^{\mathsf{T}}\Omega(\boldsymbol{u}-\Lambda\boldsymbol{w}-\boldsymbol{\lambda})\right\},$$

for conformable matrices Λ and Ω and vector $\boldsymbol{\lambda}$, dependent on θ, i.e. normal with mean and variance

$$\begin{aligned} \mathrm{E}(\boldsymbol{u}|\boldsymbol{w};\theta) &= \Lambda\boldsymbol{w}+\boldsymbol{\lambda}, \\ \mathrm{Var}(\boldsymbol{u}|\boldsymbol{w};\theta) &= \Omega^{-1}, \end{aligned} \tag{10}$$

respectively. Then, from (3) the approximation to the conditional density of $\boldsymbol{u}|\boldsymbol{y};\theta$, $\hat{f}(\boldsymbol{u}|\boldsymbol{y};\theta)$, is normal with mean and variance respectively

$$\begin{aligned} \widehat{\mathrm{E}}(\boldsymbol{u}|\boldsymbol{y};\theta) &= \Lambda\hat{\boldsymbol{w}}+\boldsymbol{\lambda} \\ \widehat{\mathrm{Var}}(\boldsymbol{u}|\boldsymbol{y};\theta) &= \Omega^{-1}+\Lambda\hat{H}^{-1}\Lambda^{\mathsf{T}}. \end{aligned} \tag{11}$$

The predictive density, $f(\boldsymbol{u}|\boldsymbol{y}) = \int f(\boldsymbol{u}|\boldsymbol{y};\theta)f(\theta|\boldsymbol{y})\,\mathrm{d}\theta$, is then approximated by a multivariate normal mixture by

$$\hat{f}(\boldsymbol{u}|\boldsymbol{y}) = \sum_{j=1}^{a} W_j \hat{f}(\boldsymbol{u}|\boldsymbol{y};\theta_j). \tag{12}$$

similarly to (8). Prediction intervals for the components of $\boldsymbol{u}$ are then computed using the appropriate quantiles of the predictive distribution (12).

Suppose further that we are interested in a function $g(u)$ where u is a single element of $\boldsymbol{u}$. For example, in the context of disease mapping, when u is the linear predictor at some location, $g(u) = e^u/(1+e^u)$ denotes the prevalence of the disease at that location, and interest lies in creating a prediction map for $g(u)$. For given data $\boldsymbol{y}$ the prediction is

$$\mathrm{E}(g(u)|\boldsymbol{y}) = \int g(u)f(u|\boldsymbol{y})\,\mathrm{d}u. \tag{13}$$

If $g(\cdot)$ is a non-linear function of u, simply replacing u by its expectation leads to biased predictions. Of course, if $g(\cdot)$ is monotone, a prediction interval can be obtained by transforming the prediction interval from (12), but for the purpose of creating a prediction map, a point prediction is needed.

The solution we propose is to evaluate (13) numerically using the ideas of Langrock (2011). To that end, let $u_0, \ldots, u_b$ be an ordered grid of values of u. These values are chosen in order to contain most of the mass of the marginal of u from approximation (12). Then

$$\begin{aligned}
\mathrm{E}(g(u)|\boldsymbol{y}) &= \int g(u) f(u|\boldsymbol{y}) \, \mathrm{d}u \\
&\approx \int g(u) \hat{f}(u|\boldsymbol{y}) \, \mathrm{d}u \\
&\approx \sum_{i=1}^{b} g\left(\frac{u_{i-1}+u_i}{2}\right) \int_{u_{i-1}}^{u_i} \hat{f}(u|\boldsymbol{y}) \, \mathrm{d}u \\
&= \sum_{i=1}^{b} g\left(\frac{u_{i-1}+u_i}{2}\right) \sum_{j=1}^{a} W_j \int_{u_{i-1}}^{u_i} \hat{f}(u|\boldsymbol{y};\theta_j) \, \mathrm{d}u, \qquad (14)
\end{aligned}$$

where the last integral is over a univariate normal density and is straightforward to compute. An approximation to the prediction variance is obtained similarly.

3.2 Application to GGLM

The distribution of the linear predictor is

$$f(\boldsymbol{w}|\beta,\theta) = (2\pi)^{-\frac{N}{2}} |\mathsf{T}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{w}-X\beta)^{\intercal}\mathsf{T}^{-1}(\boldsymbol{w}-X\beta)\right\}, \qquad (15)$$

i.e. normal with mean $X\beta$ and variance-covariance matrix $\mathsf{T} = A\boldsymbol{\Sigma}A^{\intercal} + D$.

We will distinguish two cases for the prior distribution for β, which are the most common found in the literature, the normal and flat improper priors. These are denoted respectively by

$$\beta \sim \mathrm{N}_p(\xi_0, \Phi_0^{-1}),$$

for given values ξ_0 and Φ_0, and

$$\beta \sim \mathrm{U}_p(-\infty, \infty).$$

In each of these two cases we can in fact integrate out β from (15), and obtain the distribution of the linear predictor without conditioning on the regressor coefficients. In both cases this distribution will be the normal distribution but with different mean and variance each time. We will denote by λ_N and λ_U the means, and by Υ_N and Υ_U the precision matrices when the prior for β is respectively

normal or uniform. Then

$$\begin{aligned} \lambda_N &= X\xi_0, \quad \Upsilon_N = \mathsf{T}^{-1} - \mathsf{T}^{-1}X(X^\intercal \mathsf{T}^{-1}X + \Phi_0)^{-1}X^\intercal \mathsf{T}^{-1}, \\ \lambda_U &= 0, \qquad \Upsilon_U = \mathsf{T}^{-1} - \mathsf{T}^{-1}X(X^\intercal \mathsf{T}^{-1}X)^{-1}X^\intercal \mathsf{T}^{-1}. \end{aligned} \tag{16}$$

The distribution of $\beta|\boldsymbol{w};\theta$ is also normal. Writing ξ_N, ξ_U for the means and Φ_N, Φ_U for the precision matrices in the case of the normal and uniform priors, we have

$$\begin{aligned} \xi_N &= \Phi_N^{-1}(X^\intercal \mathsf{T}^{-1}\boldsymbol{w} + \Phi_0\xi_0), \quad \Phi_N = X^\intercal \mathsf{T}^{-1}X + \Phi_0, \\ \xi_U &= \Phi_U^{-1}X^\intercal \mathsf{T}^{-1}\boldsymbol{w}, \qquad\qquad\qquad \Phi_U = X^\intercal \mathsf{T}^{-1}X. \end{aligned} \tag{17}$$

For prediction we compute the conditional mean and precision of the spatial random field given $(\boldsymbol{w};\theta)$. Let μ, P denote this mean and precision matrix. Then

$$\mu = \Sigma A^\intercal \Upsilon(\boldsymbol{w} - \lambda), \quad \mathsf{P} = \{\Sigma - \Sigma A^\intercal \Upsilon A \Sigma\}^{-1} \tag{18}$$

with λ and Υ chosen accordingly from (16). Moreover, let $\boldsymbol{z}_0$ be the value of the random field at locations Q, having mean 0 and variance-covariance matrix Σ_0, and let C denote the covariance between $\boldsymbol{z}$ and $\boldsymbol{z}_0$. Then its conditional mean and precision given $(\boldsymbol{w};\theta)$ are μ_0 and P_0 where

$$\mu_0 = C^\intercal A^\intercal \Upsilon(\boldsymbol{w} - \lambda), \quad \mathsf{P}_0 = \{\Sigma_0 - C^\intercal A^\intercal \Upsilon A C\}^{-1}. \tag{19}$$

Equations (17), (18) and (19) are of the form (10) so, given the posterior distribution for θ, their predictive distribution is obtained from (11) and (12). To that end, write

$$f(\boldsymbol{w}|\theta) \propto |\mathsf{T}|^{-\frac{1}{2}}|\Phi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{w} - \lambda)^\intercal \Upsilon(\boldsymbol{w} - \lambda)\right\}$$

for the general form of the distribution of $\boldsymbol{w}|\theta$, with Φ chosen from (17), and define

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, f(\boldsymbol{y}|\boldsymbol{w}) f(\boldsymbol{w};\theta).$$

Also let

$$\hat{H} = \Upsilon + \operatorname{diag}\{\kappa''(\hat{w}_{ij})\}.$$

Then, by Laplace approximation, the conditional distribution of $\boldsymbol{w}|\boldsymbol{y};\theta$ is approximately

$$\boldsymbol{w}|\boldsymbol{y};\theta \overset{\text{approx}}{\sim} \mathrm{N}_N(\hat{\boldsymbol{w}}, \hat{H}^{-1}),$$

and the posterior for θ becomes

$$\hat{f}(\theta|\boldsymbol{y}) \propto f(\theta) \cdot |\hat{H}|^{-\frac{1}{2}} |\mathsf{T}|^{-\frac{1}{2}} |\Phi|^{-\frac{1}{2}} \exp\left\{\boldsymbol{y}^{\mathsf{T}}\hat{\boldsymbol{w}} - \sum \kappa(\hat{w}_i)\right\} \exp\left\{-\frac{1}{2}(\hat{\boldsymbol{w}} - \lambda)^{\mathsf{T}} \Upsilon (\hat{\boldsymbol{w}} - \lambda)\right\}. \tag{20}$$

3.3 A note on the asymptotics

Central to the INLA methodology is the application of Taylor expansion on (1) which results to approximations (3) and (4). A regulatory condition for the Taylor approximation, and hence the Laplace approximation, to be valid is that the dimension of the domain of the function that is being approximated is constant. Unfortunately, this is not the case for the GGLM, where the two common asymptotic regimes *increasing-domain* and *infill* asymptotics assume that the number of sampling locations, k, increases to infinity (Stein, 1999, Section 3.3).

Shun and McCullagh (1995) studied the performance of the Laplace approximation under non-standard asymptotic settings. The case for GGLM was discussed further by Evangelou et al. (2011). In short, the validity of Laplace approximation for GGLM requires increasing domain asymptotics and the assumption that $(m_{ij}n_i)/k \to \infty$. Binary data, for example, where there is one observation per location, do not satisfy these requirements and application of the methodology in this case may lead to inconsistent results. Further research in this area is necessary for a methodology that will include these cases as well.

3.4 Full-scale approximation to the covariance

Application of the INLA methodology described in the earlier sections requires the inversion and determinant of $k \times k$ matrices. For example, for computing the inverse of the $N \times N$ matrix $\mathsf{T} = D + A\Sigma A^{\mathsf{T}}$, we may use the formula

$$\mathsf{T}^{-1} = D^{-1} - D^{-1}A\left(\Sigma^{-1} + A^{\mathsf{T}}D^{-1}A\right)^{-1}A^{\mathsf{T}}D^{-1}.$$

The computations become intractable if the dimension k is large and Σ is dense, and a big part of the literature in spatial statistics is concerned with tackling this issue.

Recently, Sang and Huang (2012) have proposed the idea of full-scale approximation to the covariance matrix. The approximation starts by specifying a set of k_L *knot* locations S_L, where k_L is much smaller than k, which cover the region of interest sufficiently. The predictive process approximation to Σ is then (Banerjee et al., 2008)

$$\Gamma \Sigma_L^{-1} \Gamma^\intercal,$$

where Σ_L is the covariance matrix at locations S_L and Γ denotes the covariance of the random field between S and S_L. The predictive process approximation is efficient in capturing the large-scale dependence structure of the random field but may not perform well for short distances. The approximation can be improved by adding to it the sparse matrix

$$\Sigma_S := (\Sigma - \Gamma \Sigma_L^{-1} \Gamma^\intercal) \odot E_\gamma,$$

a technique known as *tapering* (Furrer et al., 2006), where $\odot$ denotes the elementwise product between two matrices, and E_γ is a sparse correlation matrix on S, arising from a compactly supported correlation function (Gneiting, 2002) e.g. the spherical correlation, with range γ. Then, the full scale approximation to the matrix Σ becomes

$$\tilde{\Sigma} := \Sigma_S + \Gamma \Sigma_L^{-1} \Gamma^\intercal.$$

Once the full-scale approximation is available, it can be incorporated into the INLA methodology as outlined in Appendix B.

Working with $\tilde{\Sigma}$ is much easier computationally than Σ, as discussed by Sang and Huang (2012). In particular, the computational cost of using the original $k \times k$ covariance matrix Σ is typically to the order of $O(k^3)$ while the computational cost of using the approximation $\tilde{\Sigma}$ is to the order of $O(k(k_L^2 + k_T^2))$ where k_L is the number of knots and k_T is the average number of non-zero elements in the rows of $\tilde{\Sigma}$ which are both much smaller than k.

A number of important questions arise at this point. (a) What is the optimal choice for knot

locations? and (b) What is the optimal choice for the taper range γ? Finley et al. (2009) discuss the first question in the context of Gaussian data by putting it in the framework of spatial design (Müller, 2007). To that end, Finley et al. (2009) concluded that a sequential design, i.e. selecting the knot locations one at a time, gives better results than a space-filling design. As a design criterion they used the average prediction variance. However, it is not straightforward how to compute the prediction variance for non-Gaussian responses. On the other hand, Evangelou and Zhu (2012) showed that, if the sample size at each location is large, the optimal design for GGLM comes close to the optimal design for the Gaussian geostatistical model. With this in mind, the sequential updating scheme proposed in Finley et al. (2009), remains a good choice in GGLM as well.

In terms of choosing the taper range γ, Kaufman et al. (2008) suggest to begin with a low value and then increase it sequentially. As γ increases, the approximation error becomes smaller but the computational time increases and a choice that balances accuracy against computational speed should be made. In fact for very large values of γ there is very little improvement in terms of accuracy and speed, if at all. In this paper we measure accuracy in terms of the Frobenius distance, defined as

$$d(\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}}) := \mathrm{tr}\{(\boldsymbol{\Sigma} - \tilde{\boldsymbol{\Sigma}})^{\top}(\boldsymbol{\Sigma} - \tilde{\boldsymbol{\Sigma}})\}^{\frac{1}{2}},$$

where $\boldsymbol{\Sigma}$ denotes the exact covariance matrix for the spatial random field $\boldsymbol{z}$, and $\tilde{\boldsymbol{\Sigma}}$ its approximation. The advantage of using the Frobenius norm, as opposed to the Kullback-Leibler divergence, is that it is faster to compute when the matrices are of large dimension. In terms of computational time, we suggest using as a proxy the execution time for the Cholesky decomposition of $\tilde{\boldsymbol{\Sigma}}$. The Frobenius norm and the computational time are measured for different values of the taper range and then scaled from 0 to 1. These values are then plotted together against the taper range to produce a plot similar to the one shown in Figure 2. Examination of the figure allows us to assess the tradeoff between computing time and accuracy.

4 Examples

4.1 A simulated example

In this example the performance of INLA with and without the approximation to the covariance function is compared to MCMC.

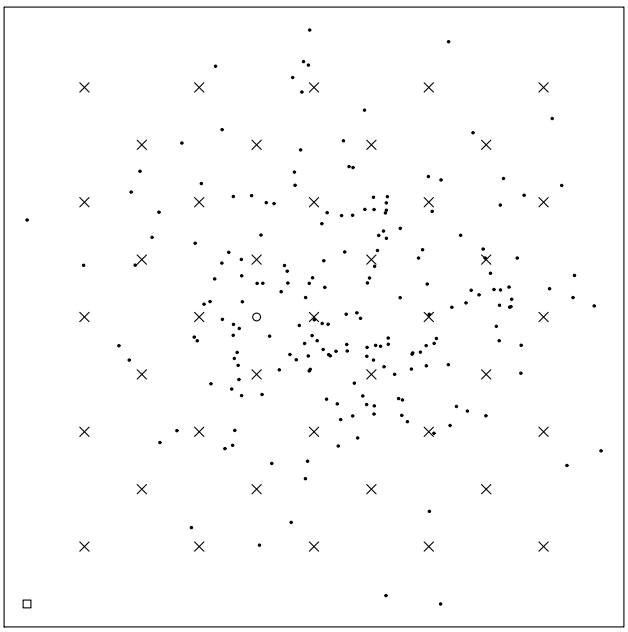

Figure 1: Locations for the simulated example, indicated by ·, and grid for the full scale approximation, indicated by ×. Prediction is considered at a central site (○) and a far site (□).

A total of $k = 200$ locations were chosen within the region $(0, 1) \times (0, 1)$ as shown in Figure 1, from where a Gaussian random field was sampled. The observations consist of a binomial GGLM with logit link and exponential covariance function, and $m = 100$ observations were drawn at each location. The linear predictor at location $\boldsymbol{x} = (x_1, x_2)$, $w_{\boldsymbol{x}}$, was set to

$$w_{\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + z_{\boldsymbol{x}} + \epsilon_{\boldsymbol{x}},$$

with $\beta_0 = -1$, $\beta_1 = \beta_2 = 1$, and covariance parameters $\sigma^2 = 0.5$, $\rho = 0.4$, and $\epsilon_{\boldsymbol{x}} \sim \mathrm{N}(0, \tau^2)$ i.i.d. with $\tau^2 = 0.1\sigma^2$.

The regressor coefficients were assigned flat improper priors, σ^2 an inverse gamma prior with parameters $a = b = 0.1$, and ρ a uniform prior in $(0, 1)$. The ratio τ^2/σ^2, called *relative nugget* by Christensen (2004) was assumed known.

The INLA method was used to estimate the parameters of the model and the random field at the observed locations. We ran the INLA method with three different covariance functions, the exact exponential covariance, its full-scale approximation with knots corresponding to a triangular grid, shown by a × in Figure 1 and using the spherical correlation for tapering, and the predictive process approximation using the same knots.

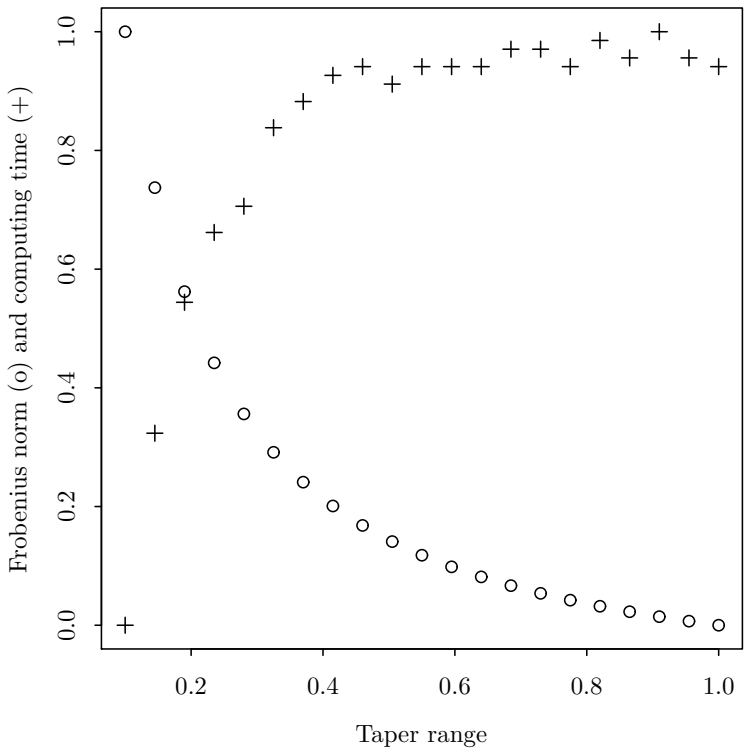

Figure 2: Scaled Frobenius norm ($\circ$) and computational time ($+$) against taper range γ.

In order to choose an appropriate value for the taper range, the approximation is computed for $\rho = 0.1, 0.2, \ldots, 1.5$ and for $\gamma = 0.100, 0.145, \ldots, 1.000$ (the maximum distance between locations is 1.078). The Frobenius norm between the exact and approximate covariance matrices, $\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Sigma}}$ respectively, was computed as well as the computational time for performing Cholesky decomposition on $\tilde{\boldsymbol{\Sigma}}$. The norm and computational time are scaled linearly in order to range from 0 to 1 and are plotted against the different values of the taper range. Figure 2 shows such plot for the case $\rho = 0.5$ (other values of ρ produce a similar plot). Based on this plot, we choose the taper range to be $\gamma = 0.2$ as this allows an approximation errror of about 40% with increase of about 40% in computing time.

The `R` package `geoRglm` (R Core Team, 2018, Christensen and Ribeiro Jr, 2002) was used for the MCMC. The length of the sample was 1,000,100, with the first 10,000 being discarded as burn-in and the remaining samples were thinned by 100, thus retaining 9901 random samples from the posterior distribution of the parameters and the random field.

The posterior densities for the two covariance parameters and the intercept using the different methods are shown in Figure 3 (the plots for the other two regressor coefficients are similar and are

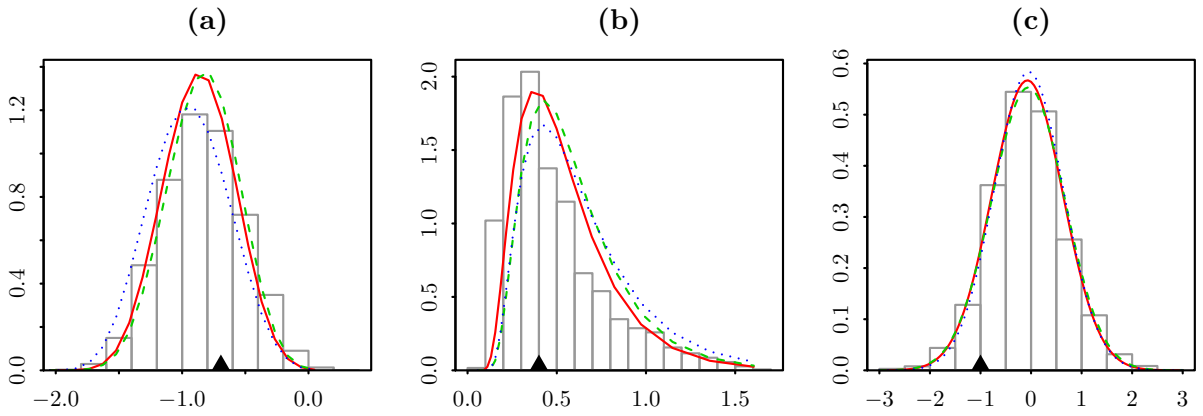

Figure 3: Posterior densities for (a) logarithm of sill, (b) range, and (c) intercept. The histogram shows the MCMC sample. The approximation using INLA and exact covariance matrix is shown by a solid line, the INLA with the full scale approximation is shown by a dashed line, and the INLA with the predictive process approximation is shown by a dotted line. The true parameter value is indicated by a triangle on the horizontal axis.

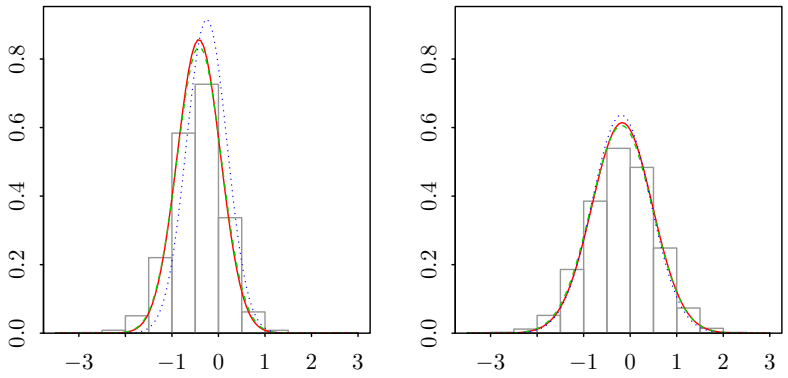

Figure 4: Predictive distribution of the random field at a central site (left) and a far site (right). The histogram shows the MCMC sample. The approximation using INLA and exact covariance matrix is shown by a solid line, the INLA with the full scale approximation is shown by a dashed line, and the INLA with the predictive process approximation is shown by a dotted line.

not shown). We observe that the INLA with the exact covariance agrees with the MCMC sample and the method using the full-scale approximation matches closely with the exact. The predictive process approximation may have a small bias when used to estimate the covariance parameters but it still works well for the estimation of the regressor coefficients.

We also considered prediction at two sites, a central site, indicated by a circle in Figure 1, and a far site, indicated by a square. The predictive distributions using each method are shown

in Figure 4. The plots show that the exact and full-scale approximations agree with the MCMC sample but the predictive process approximation exhibits small bias.

In summary, the example shows that INLA provides a good approximation to the predictive distributions of the parameters and the random field and the full-scale approximation can be used as a computationally milder alternative.

4.2 Loa loa prevalence in Cameroon

The *loa loa* is a worm-like parasite infecting human by travelling through their tissue. The young larvae develop in flies, commonly found in Africa and India, which infect human by biting them. The main methods of diagnosis include the presence of the parasite in the blood or in the eye, and the presence of skin swellings. When found, the worm is removed by a surgery. Diggle et al. (2007) discuss about the importance of estimating the prevalence of *loa loa* for precautionary reasons.

The data consist of samples from $k = 197$ villages in Cameroon and its neighbouring country Nigeria, as shown in Figure 7. For the ith village the number of infected subjects y_i and the number of tested subjects m_i were recorded. We also have information about the altitude and a measure of vegetation, the normalised-difference vegetation index (NDVI) for each month for the year 1996. Note that the individuals in this example are the villages.

The $p = 6$ explanatory variables are as in Diggle et al. (2007), although our values are slightly different from theirs. They consist of an intercept (β_0), a piecewise linear effect of the altitude with the regression coefficient changing at .65Km and 1Km (β_1, β_2, β_3), a linear effect of the maximum NDVI when it is below the 0.8 level β_4, and a linear effect of the standard deviation of NDVI β_5. We assign flat improper priors for all regressor coefficients.

Based on the above information, an appropriate model for the linear predictor is

$$\boldsymbol{w} = X\beta + \boldsymbol{z} + \boldsymbol{\epsilon}.$$

The number of infected individuals in the ith village conditioned on the linear predictor is modelled as binomial GGLM with logit link, i.e.

$$y_i|w_i \sim \text{Bin}\left(m_i, \frac{e^{w_i}}{1+e^{w_i}}\right).$$

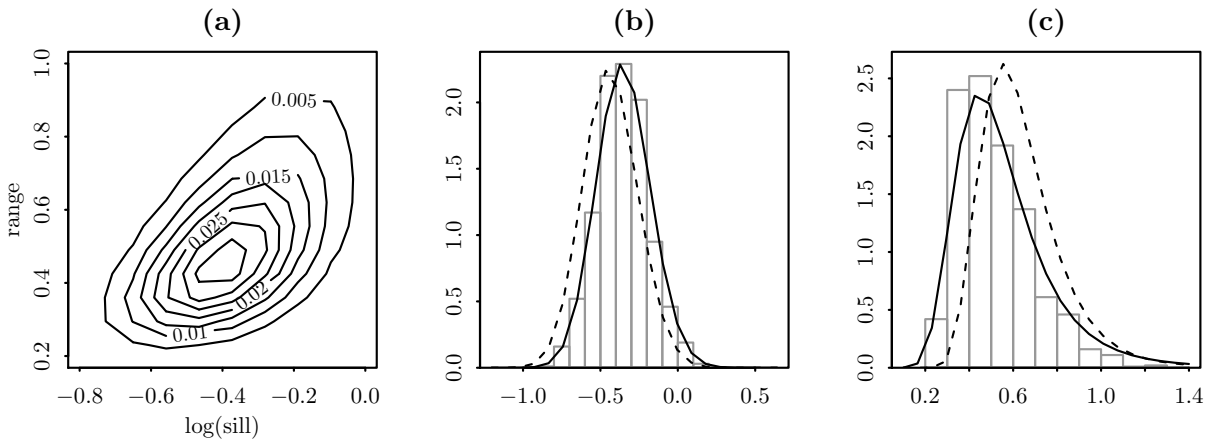

Figure 5: Posterior plots for the variance parameters. (a) Joint posterior of $\log(\sigma^2)$ and ρ using exact INLA, (b) Marginal posterior of $\log(\sigma^2)$, (c) Marginal posterior of ρ. The histogram is for the MCMC sample, the exact INLA is shown by a solid line and the full-scale INLA by a dadhed line.

	Exact INLA			Full-scale INLA			MCMC		
Parameter	Estimate	95% interval		Estimate	95% interval		Estimate	95% interval	
Intercept (β_0)	−14.17	−18.58	−9.76	−15.03	−19.28	−10.77	−14.67	−19.16	−9.90
Elevation $0-.65$Km (β_1)	2.28	1.07	3.49	2.19	1.02	3.36	2.35	1.15	3.60
Elevation $.65-1$Km (β_2)	1.62	0.90	2.34	1.60	0.91	2.29	1.68	1.00	2.39
Elevation $1-1.3$Km (β_3)	0.81	0.17	1.45	0.68	0.05	1.30	0.83	0.18	1.46
Max(NDVI) (β_4)	14.09	8.00	20.17	15.11	9.16	21.06	14.66	8.19	20.82
Sd(NDVI) (β_5)	0.71	−9.68	11.10	1.27	−8.87	11.42	0.68	−9.04	11.22
Sill (σ^2)	0.72	0.50	1.02	0.66	0.45	0.94	0.70	0.51	0.99
Range (ρ)	0.55	0.25	1.08	0.64	0.36	1.08	0.48	0.28	0.92

Table 1: Parameter estimates for the loa loa prevalence in Cameroon using exact and approximate INLA.

Each of the regressor coefficients is assumed to have an independent improper flat prior.

The random field $\boldsymbol{z}$ is assumed to have an exponential covariance structure. In addition, the relative nugget parameter is fixed to $\tau^2/\sigma^2 = 0.4$ as in Diggle et al. (2007), and we assume a uniform improper prior for σ^2, i.e. $\sigma^2 \sim \mathrm{U}(0, \infty)$, while for the range parameter we assign $\rho \sim \mathrm{U}(0.1, 1.4)$.

We consider model fitting using INLA with exact and approximate covariance matrix at 36 knots forming a square grid. Figure 5 shows the posterior density of $(\log \sigma^2, \rho)|\boldsymbol{y}$, evaluated at 21×21 equidistant points for each parameter using the two methods. The two methods agree in terms of $\log \sigma^2$ but some discrepancies exist when estimating ρ. There are also small discrepancies for some of the regressor coefficients as can be seen in Figure 6 which suggests that a larger number of knots is required.

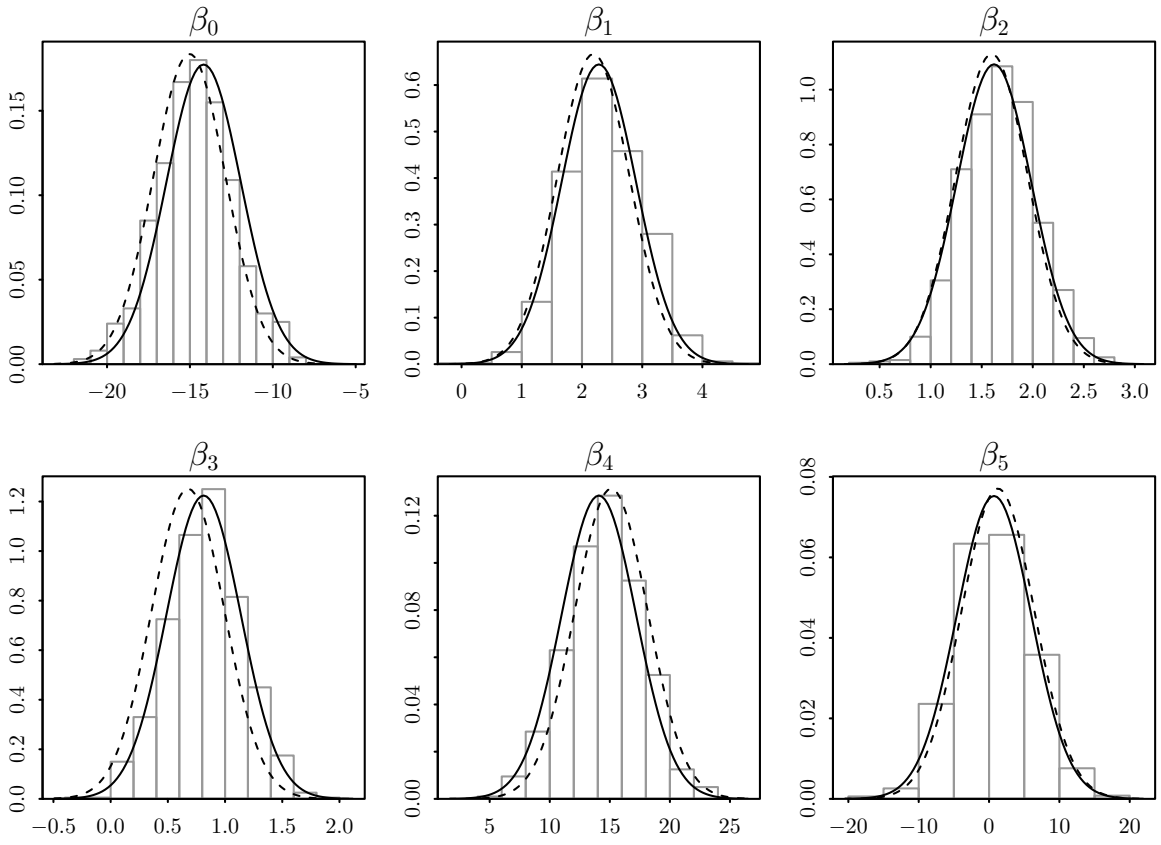

Figure 6: Posterior for the regressor coefficients. The histogram is for the MCMC sample, the exact INLA is shown by a solid line and the full-scale INLA by a dashed line.

A 95% confidence interval for the variance parameters is obtained by spline approximation to the marginal posterior for each parameter. For the regressor coefficients the confidence interval is obtained assuming normality of the estimator. These results are summarised in Table 1. Besides the standard deviation of NDVI, all other coefficients are found to be significant using both methods.

A probability map of the prevalence of *loa loa* is of interest. Let $w = \boldsymbol{x}^{\mathsf{T}}\beta + z + \epsilon$ denote the linear predictor at any given location. The prevalence is then given by

$$\text{prevalence} = \frac{e^w}{1 + e^w},$$

and is estimated using (14). Figure 7 shows the predicted prevalence and prediction standard deviation for the region of interest. The map closely resembles the one reported by Diggle et al. (2007). We also note that the uncertainty of the prediction is higher where there are no sampled

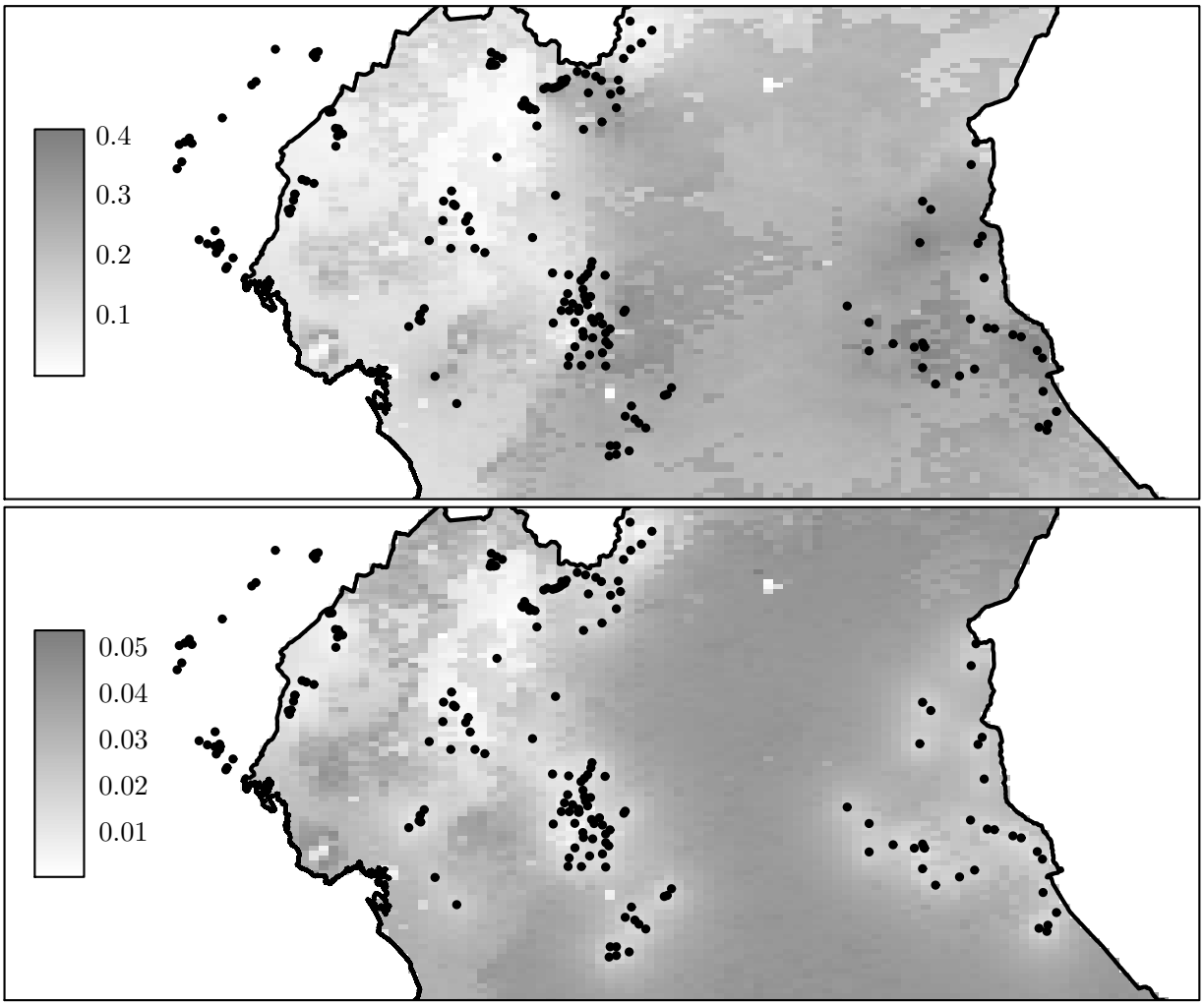

Figure 7: Predicted prevalence of the *loa loa* parasite (top), and prediction standard deviation (bottom).

villages close by.

4.3 Childhood malaria in the Gambia

In this study, presented by Diggle et al. (2002), the data are obtained by sampling children in $k = 65$ villages in the Gambia as shown in Figure 8. These villages are classified into 5 areas (respectively, south-west, north-west, centre, north-east, and south-east), and 42 of them belong to the primary health care structure (PHC) of the Ministry of Health, while the remaining 23 do not.

For the jth child in the ith village an indicator y_{ij} of the presence of malaria in a blood sample was recorded, along with the child's age, whether or not it regularly slept under a bed net, and if

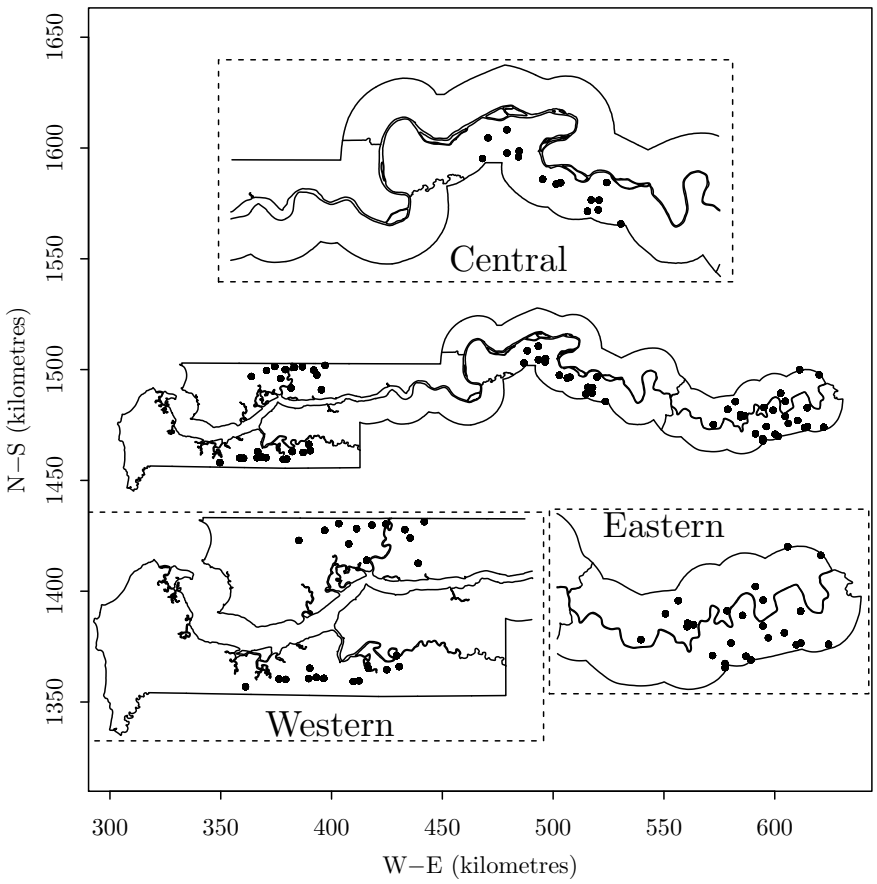

Figure 8: Sampled locations for the Gambia data from Ribeiro Jr and Diggle (2001).

so whether the net was treated with insecticide.

The $p = 10$ explanatory variables for each individual consist of the intercept, age (in years), a three-level effect for bed net (no net, untreated, treated), the amount of greenness at the corresponding village, an indicator of whether the village belongs to the PHC, and a five-level effect for the area. Note that some of these variables correspond to the individual within the village so the response variable is the indicator y_{ij} of whether the jth child in the ith village is infected or not. Let $\boldsymbol{x}_{ij}$ denote the set of explanatory variables for the (i, j) observation. The linear predictor is then given by

$$w_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + z_i + \epsilon_{ij}.$$

We adopt the binomial GGLM with logit link, i.e.

$$y_{ij}|w_{ij} \sim \text{Bin}\left(1, \frac{e^{w_{ij}}}{1 + e^{w_{ij}}}\right).$$

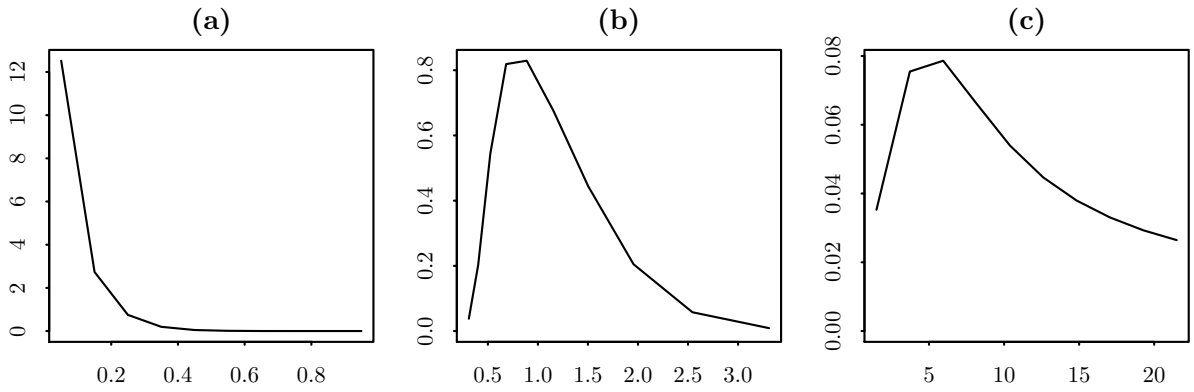

Figure 9: Posterior densities for the parameters (a) τ^2, (b) σ^2, and (c) ρ of the Gambia malaria data.

Parameter	Estimate	95% interval	
Intercept (β_0)	−0.07309	−2.95100	2.80483
Age (β_1)	0.00066	0.00042	0.00090
Untreated bed net (β_2)	−0.36216	−0.67639	−0.04793
Treated bed net (β_3)	−0.68297	−1.07497	−0.29097
Greenness (β_4)	−0.01334	−0.07507	0.04839
PHC (β_5)	−0.32790	−0.77921	0.12340
Area 2 (β_6)	−0.69385	−2.26728	0.87958
Area 3 (β_7)	−0.78240	−2.44258	0.87778
Area 4 (β_8)	0.65537	−1.12152	2.43226
Area 5 (β_9)	0.97627	−0.80963	2.76217
Nugget (τ^2)	0.13209	0.00310	0.26136
Sill (σ^2)	0.98459	0.34501	1.82461
Range (ρ)	9.82025	0.54713	18.63800

Table 2: Parameter estimates of the Gambia malaria data.

For covariance, we use the exponential model. The prior distributions for the covariance parameters are chosen to be

$$\tau^2 \sim \text{IG}(0.01, 0.01), \quad \sigma^2 \sim \text{IG}(0.01, 0.01), \quad \rho \sim \text{U}(1.5, 21.5),$$

i.e. inverse gamma, inverse gamma, and uniform.

The posterior densities for these parameters are shown in Figure 9. Table 2 displays a summary of the parameter estimates and a 95% confidence interval for each parameter.

We consider prediction of the spatial random field in the region which is useful in identifying high-risk areas. The prediction, and its standard deviation are shown in Figure 10. The sampled

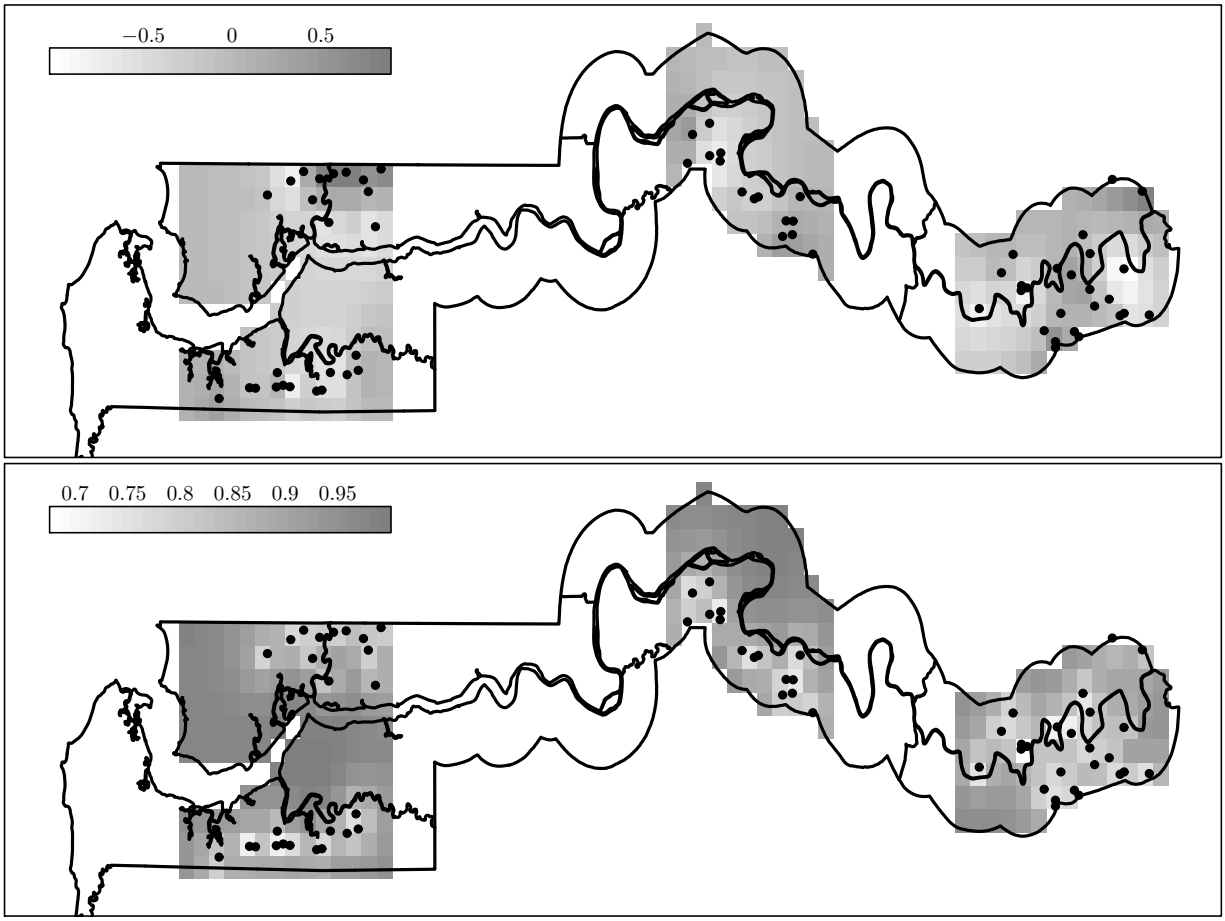

Figure 10: Prediction of spatial random field for the Gambia malaria data (top) and prediction standard deviation (bottom).

locations are also indicated. Prediction far from the sampled sites is associated with high degree of uncertainty so it is not attempted. The pattern is similar to the one reported by Diggle et al. (2002).

5 Conclusion

In this paper a methodology for approximate inference in Bayesian geostatistical models is presented. The application of full scale approximation to the spatial covariance matrix is discussed and its implementation within INLA is demonstrated. Issues regarding the choice of parameters for the full-scale approximation are addressed. In terms of choosing the knots for the predictive process approximation we use spatial design for GGLM while for choosing the taper range, a graphical method is presented. Furthermore, we discuss prediction of transformed parameters using the

INLA method.

The proposed method is compared against MCMC in a simulated example and found to have good performance. The only caveat is that INLA with the full scale approximation may exhibit bias when estimating the posterior distribution for the range parameter, therefore more research to this direction is needed. Despite this, it is evident from the two examples that our method can be a useful tool for geostatistical inference in GGLM.

The proposed model is especially useful for disease mapping as demonstrated by the examples. Its potential for extending it to allow for autoregressive errors, in order to account for temporal effects, is left for future research.

A Selecting grid values for the variance parameters

In this section we explain how we obtain a grid of values $\{\theta_1, \ldots, \theta_a\}$ for the discrete approximation to the posterior $f(\theta|\boldsymbol{y})$ when the number a is pre-specified. Since the elements of θ are positive, in practice we work with the logarithmic transformation of its elements, however in the following we don't make a distinction on the scale used.

The suggestion of Rue et al. (2009) is to create a grid within a confidence interval (of some high level, say 99%) obtained by maximising (20). In our case, the objective function for the maximisation is the logarithm of (20), i.e.

$$h(\theta|\boldsymbol{y}) = \log f(\theta) - \frac{1}{2}\log|\hat{H}| - \frac{1}{2}\log|\mathsf{T}| - \frac{1}{2}\log|\Phi| + \boldsymbol{y}^{\mathsf{T}}\hat{\boldsymbol{w}} - \sum \kappa(\hat{w}_i) - \frac{1}{2}(\hat{\boldsymbol{w}} - \lambda)^{\mathsf{T}}\Upsilon(\hat{\boldsymbol{w}} - \lambda),$$

with $\hat{\boldsymbol{w}}$ chosen to satisfy

$$\boldsymbol{y} - \hat{\boldsymbol{\kappa}}' - \Upsilon(\hat{\boldsymbol{w}} - \lambda) = 0, \tag{21}$$

and $\hat{\boldsymbol{\kappa}}'$ denotes the vector with ith element $\kappa'(\hat{w}_i)$, $i = 1, \ldots, N$. Differentiating (21) with respect to θ_j we obtain

$$\partial_j \hat{\boldsymbol{w}} = -\hat{H}^{-1}\left(\partial_j \Upsilon\right)(\hat{\boldsymbol{w}} - \lambda),$$

where $\partial_j = \frac{\partial}{\partial \theta_j}$.

The first order derivative of the objective function is

$$\begin{aligned}\partial_j h(\theta|\boldsymbol{y}) &= \frac{\partial_j f(\theta)}{f(\theta)} - \frac{1}{2}\operatorname{tr}(\hat{H}^{-1}\partial_j \hat{H}) - \frac{1}{2}\operatorname{tr}(\mathsf{T}^{-1}\partial_j \mathsf{T}) - \frac{1}{2}\operatorname{tr}(\Phi^{-1}\partial_j \Phi) \\ &\quad + \boldsymbol{y}^\intercal \partial_j \hat{\boldsymbol{w}} - \hat{\boldsymbol{\kappa}}'^\intercal \partial_j \hat{\boldsymbol{w}} - \frac{1}{2}(\hat{\boldsymbol{w}} - \lambda)^\intercal (\partial_j \Upsilon)(\hat{\boldsymbol{w}} - \lambda) - (\hat{\boldsymbol{w}} - \lambda)^\intercal \Upsilon \partial_j \hat{\boldsymbol{w}} \\ &= \frac{\partial_j f(\theta)}{f(\theta)} - \frac{1}{2}\operatorname{tr}(\hat{H}^{-1}\partial_j \hat{H}) - \frac{1}{2}\operatorname{tr}(\mathsf{T}^{-1}\partial_j \mathsf{T}) - \frac{1}{2}\operatorname{tr}(\Phi^{-1}\partial_j \Phi) - \frac{1}{2}(\hat{\boldsymbol{w}} - \lambda)^\intercal (\partial_j \Upsilon)(\hat{\boldsymbol{w}} - \lambda),\end{aligned}$$

simplified using (21), which can be used in a quasi-Newton iteration scheme for maximising the posterior for θ. The negative inverse Hessian at the maximum is used in the place of the variance and the desired confidence interval is obtained assuming normality.

B Matrix computations

Key formulae in the derivations that follow for the inversion and determinant of matrices of the form

$$\Delta + A^\intercal S A$$

are

$$\begin{aligned}(\Delta + A^\intercal S A)^{-1} &= \Delta^{-1} - \Delta^{-1} A^\intercal (S^{-1} + A\Delta^{-1}A^\intercal)^{-1} A\Delta^{-1}, \\ |\Delta + A^\intercal S A| &= |\Delta||S||S^{-1} - A\Delta^{-1}A^\intercal|.\end{aligned} \tag{22}$$

Below we show how these results can be used to approximate the inverse of the approximate covariance matrix. Similar techniques are used for computing its determinant.

Given the approximation to the variance-covariance matrix Σ

$$\tilde{\Sigma} = \Sigma_S + \Gamma \Sigma_L^{-1} \Gamma^\intercal,$$

the variance covariance matrix of the linear predictor $\boldsymbol{w}$ is

$$\mathsf{T} = D + A\Sigma A^\intercal = D + A\Sigma_S A^\intercal + A\Gamma\Sigma_L^{-1}\Gamma^\intercal A^\intercal$$

with inverse

$$\mathsf{T}^{-1} = (D + A\Sigma_S A^\intercal)^{-1} - (D + A\Sigma_S A^\intercal)^{-1} A\Gamma \left\{\Sigma_L + \Gamma^\intercal A^\intercal (D + A\Sigma_S A^\intercal)^{-1} A\Gamma\right\}^{-1} \Gamma^\intercal A^\intercal (D + A\Sigma_S A^\intercal)^{-1}.$$

The matrix inside the curly brackets is of low dimension, and its inverse is available. The matrix $D + A\Sigma_S A^\intercal$ is N-dimensional, and, even though it may be sparse, computing its inverse directly may not be the best approach due to its high dimension. Using the key formula (22) for the matrix inverse,

$$(D + A\Sigma_S A^\intercal)^{-1} = D^{-1} - D^{-1}A(\Sigma_S^{-1} + A^\intercal D^{-1} A)^{-1} A^\intercal D^{-1}, \tag{23}$$

and noting that the matrix in parentheses is now k-dimensional and sparse, since it consists of a sum of two sparse matrices, can significantly ease computations.

In a similar manner we compute

$$\begin{aligned}\hat{H}^{-1} &= (K'' + \Upsilon)^{-1} \\ &= (K'' + \mathsf{T}^{-1})^{-1} + (K'' + \mathsf{T}^{-1})^{-1}\mathsf{T}^{-1} X \left\{\Phi_0 + X^\intercal (K''^{-1} + \mathsf{T})^{-1} X\right\}^{-1} X^\intercal \mathsf{T}^{-1} (K'' + \mathsf{T}^{-1})^{-1}\end{aligned}$$

with

$$\begin{aligned}(K''^{-1} + \mathsf{T})^{-1} = (K''^{-1} + D + A\Sigma_S A^\intercal)^{-1} &- (K''^{-1} + D + A\Sigma_S A^\intercal)^{-1} A\Gamma \\ &\cdot \left\{\Sigma_L + \Gamma^\intercal A^\intercal (K''^{-1} + D + A\Sigma_S A^\intercal)^{-1} A\Gamma\right\}^{-1} \Gamma^\intercal A^\intercal (K''^{-1} + D + A\Sigma_S A^\intercal)^{-1} \\ (K'' + \mathsf{T}^{-1})^{-1} &= \mathsf{T} - \mathsf{T}(K''^{-1} + \mathsf{T})^{-1}\mathsf{T}\end{aligned}$$

and $(K''^{-1} + D + A\Sigma_S A^\intercal)^{-1}$ is computed as in (23) with D replaced by $K''^{-1} + D$.

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall Ltd.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, 13(3):702–718.

Christensen, O. F., Møller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and langevin-type markov chain monte carlo. Technical report, Department of Mathematical Sciences, Aalborg University.

Christensen, O. F. and Ribeiro Jr, P. J. (2002). geoRglm: A package for generalised linear spatial models. *R News*, 2(2):26–28.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58(2):280–286.

Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M. (2002). Childhood malaria in the gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M., and Molyneux, D. H. (2007). Spatial modelling and the prediction of loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology*, 101(6):499–509.

Eidsvik, J., Finley, A. O., Banerjee, S., and Rue, H. (2012). Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics & Data Analysis*, 56(6):1362–1380.

Eidsvik, J., Martino, S., and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian journal of statistics*, 36(1):1–22.

Evangelou, E. and Maroulas, V. (2017). Sequential empirical Bayes method for filtering dynamic spatiotemporal processes. *Spatial Statistics*, 21(A):114–129.

Evangelou, E. and Zhu, Z. (2012). Optimal predictive design augmentation for spatial generalised linear mixed models. *Journal of Statistical Planning and Inference*, 142(12):3242–3253.

Evangelou, E., Zhu, Z., and Smith, R. L. (2011). Estimation and prediction for spatial generalized linear mixed models using high order laplace approximation. *Journal of Statistical Planning and Inference*, 141(11):3564–3577.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508.

Hosseini, F., Eidsvik, J., and Mohammadzadeh, M. (2011). Approximate bayesian inference in spatial glmm with skew normal latent variables. *Computational Statistics & Data Analysis*, 55(4):1791–1806.

Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, 6(4):1499–1530.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Langrock, R. (2011). Some applications of nonlinear and non-Gaussian state–space modelling by means of hidden Markov models. *Journal of Applied Statistics*, 38(12):2955–2970.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3):514–528.

McCullagh, P. and Nelder, J. A. (1999). *Generalized Linear Models*. Chapman & Hall/CRC.

Müller, W. (2007). *Collecting spatial data: optimum design of experiments for random fields*. Springer Verlag.

Paul, M., Riebler, A., Bachmann, L. M., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in medicine*, 29(12):1325–1339.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr, P. J. and Diggle, P. J. (2001). geoR: A package for geostatistical analysis. *R News*, 1(2):15–18.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Monographs on statistics and applied probability. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

Sang, H., Jun, M., and Huang, J. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics*, 5(4):2519–2548.

Schrödle, B. and Held, L. (2011). A primer on disease mapping and ecological regression using INLA. *Computational Statistics*, 26(2):241–258.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B, Methodological*, 57:749–760.

Simpson, D., Lindgren, F., and Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.

Taylor, B. M. and Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284.

Zhang, H. (2004). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.